Lecture Outline

- Gradient Projection Algorithm
- Constant Step Length, Varying Step Length, Diminishing Step Length
- Complexity Issues
- Gradient Projection With Exploration
- Projection
- Solving QPs: active set method and ADMM
- Approximating the constraint set

You should be able to . . .

- Recognise and formulate a gradient projection algorithm;
- Select the step length
- Extend the idea to exploratory moves
- Have an understanding of the complexity of the method
- Compute and approximate projections

Gradient Projection

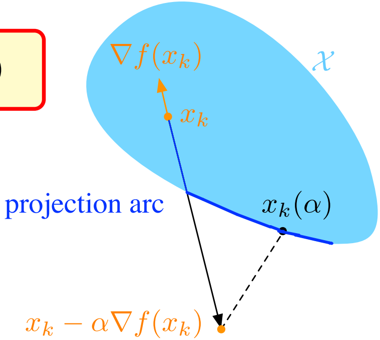
$$\min \quad f(x), \quad \text{s.t.} \quad x \in \mathcal{X}$$

- f is continuously differentiable and $\mathcal{X}$ is closed and convex.
- We have the following iteratative method:

$$x_{k+1} = \mathbf{P}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k))$$

- where $\alpha_k > 0$ is the step-length.
- Define the projection arc for $\alpha > 0$:

$$x_k(\alpha) = \mathbf{P}_{\mathcal{X}}(x_k - \alpha \nabla f(x_k))$$



Gradient Projection

- The projection arc is the set of all possible next iterates paramterised by α.
- Next we show that unless $x_k(\alpha) = x_k$ (which is a condition for optimality of x_k), the vector $x_k(\alpha) - x_k$ is a feasible descent direction.
- We first need an important result.

Theorem (Projection Theorem): *Let $\mathcal{X}$ be a nonempty closed convex subset of $\mathbb{R}^n$. There is a unique vector that minimises $\|z - x\|$ over $x \in \mathcal{X}$ called the projection of z on $\mathcal{X}$. Furthermore, $x^\star$ is the projection of z on $\mathcal{X}$ iff*

$$(z - x^\star)^T (x - x^\star) \leq 0, \quad \forall x \in \mathcal{X}.$$

Gradient Projection

Theorem (Descent Properties of Gradient Projection):

(i) *If* $x_k(\alpha) \neq x_k$*, then* $x_k(\alpha) - x_k$ *is a feasible descent direction and particularly*

$$\nabla f(x_k)^T (x_k(\alpha) - x_k) \leq -\frac{1}{\alpha} \|x_k(\alpha) - x_k\|^2, \quad \forall \alpha > 0$$

(ii) *If* $x_k(\alpha) = x_k$ *for some* $\alpha > 0$ *then* x_k *satisfies the necessary condition for minimising* $f(x)$ *over* X*, i.e.*

$$\nabla f(x_k)^T (x - x_k) \geq 0, \quad \forall x \in \mathcal{X}$$

- From Projection Theorem:

$$(x_k - \alpha \nabla f(x_k) - x_k(\alpha))^T (x - x_k(\alpha)) \leq 0, \quad \forall x \in \mathcal{X}$$

- Setting $x = x_k$ yields (i). If $x_k(\alpha) = x_k$ for some $\alpha > 0$. Above inequality yields (ii). □

Gradient Projection

- An (often) important assumption for constant step-size convergence: Lipschitz continuity of the gradient

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{X}$$

- This condition results in the following important inequality:

$$f(y) \leq \overbrace{f(x) + \nabla f(x)^T(y - x)}^{\ell(y;x)} + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{X}$$

$$f(y) - f(x) = \int_0^1 (y - x)^T \nabla f(x + t(y - x))dt$$

$$\leq \int_0^1 (y - x)^T \nabla f(x)dt + \left| \int_0^1 (y - x)^T (\nabla f(x + t(y - x)) - \nabla f(x))dt \right|$$

$$\leq (y - x)^T \nabla f(x) + \int_0^1 \|y - x\| \|\nabla f(x + t(y - x)) - \nabla f(x)\| dt$$

$$\leq (y - x)^T \nabla f(x) + \|y - x\| \int_0^1 \|y - x\| Ltdt = (y - x)^T \nabla f(x) + \frac{L}{2}\|y - x\|^2$$

Gradient Projection: Constant Step Length

Theorem (Constant Step Length Convergence): *Assume the gradient is Lipschitz continuous and $\alpha_k = \alpha$, $\alpha \in (0, 2/L)$. Then every limit point $\bar{x}$ of the generated sequence $\{x_k\}$ satisfies the necessary optimality condition*

$$\nabla f(\bar{x})^T (x - \bar{x}) \geq 0, \quad \forall x \in \mathcal{X}.$$

- From the inequality of the previous page by $y = x_{k+1}$ and $x = x_k$:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$$

- Moreover, $\nabla f(x_k)^T (x_{k+1} - x_k) \leq -\frac{1}{\alpha} \|x_{k+1} - x_k\|^2$. Thus,

$$f(x_{k+1}) \leq f(x_k) - \left(\frac{1}{\alpha} - \frac{L}{2} \right) \|x_{k+1} - x_k\|^2$$

- Since $\alpha \in (0, 2/L)$, the cost function is reduced.

Gradient Projection: Varying Step Length

- Thus for any limit $\bar{x}$ of the subsequence $\mathcal{K}$, $f(x_k) \downarrow f(\bar{x})$ and consequently $\|x_{k+1} - x_k\| \to 0$.
- Consequently,

$$\mathbf{P}_{\mathcal{X}}(\bar{x} - \alpha \nabla f(\bar{x})) - \bar{x} = \lim_{k \to \infty, k \in \mathcal{K}} x_{k+1} - x_k = 0$$

- This (by the earlier descent result) implies that $\bar{x}$ satisfies the necessary optimality condition. □

Theorem (Convergence for Convex Cost Function):
Let $\alpha_k \downarrow \bar{\alpha}$ is selected via any step length rule and for all k

$$f(x_{k+1}) \leq \ell(x_{k+1}; x_k) + \frac{1}{2\alpha_k} \|x_{k+1} - x_k\|^2.$$

Then $\{x_k\}$ converges to $x^\star$ and

$$f(x_k) - f^\star \leq \frac{\min_{x^\star \in \mathcal{X}^\star} \|x_0 - x^\star\|^2}{2k\bar{\alpha}}, \quad k \geq 0$$

Gradient Projection: Constant Step Length and Strong Convexity

- The Lipschitz condition needs to be satisfied for the level set $\mathcal{L} = \{x \in \mathcal{X} | f(x) \leq f(x_0)\}$ that depends on the initial condition.
- The error converges to 0 with an order $O(1/k)$.
- The convergence is linear when f is strongly convex.

Theorem (Convergence for Strongly Convex Cost Function): *Let $\alpha \in (0, 2/L)$ and f is strongly convex with modulus σ. Then,*

$$\|x_{k+1} - x^\star\| \leq \max(|1 - \alpha L|, |1 - \alpha\sigma|)\|x_k - x^\star\|.$$

- The bound is minimised if $\alpha = 2/(\sigma + L)$[1].
- L/σ is the condition number of the problem ($L \geq \sigma$).

[1] I have spent some part of my research finding such optimum step-lengths for different optimisation problems.

Gradient Projection: Diminishing Step Length

- Consider diminishing step size:

$$\lim_{k\to\infty} \alpha_k = 0,\ \sum_{k=0}^{\infty} \alpha_k = \infty,\ \sum_{k=0}^{\infty} \alpha_k^2 < \infty$$

- If there is a scalar γ such that

$$\gamma^2 \left(1 + \min_{x^\star \in \mathcal{X}^\star} \|x_0 - x^\star\|^2\right) \geq \sup_{k\geq 0} \|\nabla f(x_k)\|^2$$

- Then the gradient projection method converges even without Lipschitz continuity of the gradient.
- For example, $f(x) = |x|^{3/2}$ gradient projection converges to 0 for the diminishing step size, but not with a constant step size (gradient not Lipschitz)
- Convergence rate is sublinear.

Gradient Projection: Step length via an Armijo-like rule

Algorithm: Gradient Projection Via an Armijo-like rule

$\beta \in (0,1)$, x_0, α, $k \leftarrow 0$
while $\|x_{k+1} - x_k\| > \tau$ **do** ▷ e.g. or any other termination condition
  $d_k \leftarrow \mathbf{P}_{\mathcal{X}}(x_k - \alpha \nabla f(x_k)) - x_k$
  $m_k \leftarrow 0$
  while $f(x_k) - f(x_k + \beta^{m_k} d_k) < -\beta^{m_k} \nabla f(x_k)^T d_k$ **do**
    $m_k \leftarrow m_k + 1$
  end while
  $x_{k+1} \leftarrow x_k + \beta^{m_k} d_k$
  $k \leftarrow k + 1$
end while

- At each step, we search along the line $\{x_k + \gamma d_k | \gamma > 0\}$ by checking step sizes $\gamma = 1, \beta, \beta^2, \ldots$ until sufficient decrease is obtained.
- For convex f this algorithm converges to the solution without the gradient Lipschitz condition.

Some Complexity Discussions

- How many iterations are required to achieve a solution with cost that is within $\epsilon > 0$ of the optimum?
- A method has *iteration complexity* $O(1/\epsilon^p)$ if we can show (for some $M, p > 0$):

$$\min_{k \le M/\epsilon^p} f(x_k) \le f^\star + \epsilon$$

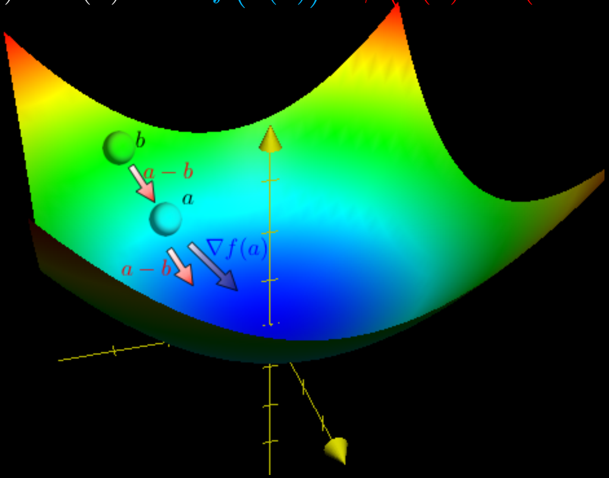
- A method involves *cost function error of order* $O(1/k^q)$ if we can show (for some $M, q > 0$):

$$\min_{j \le k} f(x_j) \le f^\star + \frac{M}{k^q}$$

- If M does not depend on n then it is good for large problems. (gradient vs. Newton)
- For gradient methods it requires $k \ge O(1/\epsilon)$ to get an error order of $O(1/k)$.
- However, these bounds do not take advantage of the special structure of the problem.

Gradient Projection With Exploration: Heavy Ball

$$x(k+1) = x(k) - \alpha \nabla f\big(x(k)\big) + \beta\big(x(k) - x(k-1)\big)$$



Gradient Projection With Exploration: Optimal Iteration Complexity

- Heavy ball takes advantage of memory (x_{k-1}) to improve the performance. Adding more memory is not necessarily useful.
- Assume $f(x)$ is convex and has a Lipschitz continuous gradient.
- The iterations become ($x_{-1} = x_0$, $\beta_k \in (0,1)$)

$$\begin{aligned} y_k &= x_k + \beta_k(x_k - x_{k-1}), \quad \text{(exploration step)} \\ x_{k+1} &= \mathbf{P}_{\mathcal{X}}(y_k - \alpha \nabla f(x_k)), \quad \text{(gradient projection step)} \end{aligned}$$

$$\beta_k = \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}.$$

- $\{\theta_k\}$ such that $\theta_0 = \theta_1 \in (0,1]$ and

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}$$

Gradient Projection With Exploration: Optimal Iteration Complexity

Example: β_k and θ_k

$$\beta_k = \begin{cases} 0 & k = 0 \\ \dfrac{k-1}{k+2} & k \geq 1 \end{cases}, \qquad \theta_k = \begin{cases} 1 & k = -1 \\ \dfrac{2}{k+2} & k \geq 0 \end{cases}$$

Theorem: *Let $\alpha = 1/L$ and β_k is chosen as above. Then,* $\lim_{k\to\infty} \|x_k - x^\star\| = 0$ *and*

$$f(x_k) - f^\star \leq \frac{2L}{(k+1)^2}\|x_0 - x^\star\|^2.$$

- The error is $O(1/k^2)$ and equivalently the iteration complexity is $O(1/\sqrt{\epsilon})$.

Gradient Projection: Projection,...what projection?

- Recall $\mathbf{P}_{\mathcal{X}}(x) = \xi$ where

$$\xi \in \arg\min_{z} \quad \|x - z\|, \quad \text{s.t.} \quad z \in \mathcal{X}$$

- So, do we need to solve another optimisation problem for each iteration of an optimisation problem?!

Example: Box Constraints

- A simple *box constraint* :

$$\mathcal{X} = \{x | l \leq x \leq u\}$$

- u and l are respectively the upper and lower bounds on the entries of x.

$$\xi_i = \begin{cases} l_i & x_i < l_i \\ x_i & l_i \leq x_i \leq u_i \\ u_i & u_i \leq x_i \end{cases}$$

Gradient Projection: What Projection?

Example: Linear Subspace $Ax = b$

- $\mathbf{P}_{\mathcal{X}}(x) = \xi$ where

$$\xi \in \arg\min_z \quad \|x - z\|, \quad \text{s.t.} \quad z \in \mathcal{X} = \{x | Ax = b\}$$

$$\xi = (I - A^T(AA^T)^{-1}A)x + A^T(AA^T)^{-1}b$$

Example: Constraint Set Defined by Inequalities

- $\mathbf{P}_{\mathcal{X}}(x) = \xi$ where

$$\xi \in \arg\min_z \quad \|x - z\|, \quad \text{s.t.} \quad z \in \mathcal{X} = \{x | c_i(x) \geq 0, i \in \mathcal{I}\}$$

- c_i concave and $\mathcal{I}$ inequality constraints index set

- This in effect is solving a quadratic problem with constraints.

Solving Quadratic Programs

- We will consider the case where the constraints $c_i(x)$ are linear.
- Two different approaches to solving QPs will be considered.
  - Primal Active Set Method
  - Alternating Direction Method of Multipliers (ADMM)
- In primal active-set methods some of the inequality constraints (and all the equalities, if any) are imposed as equalities.
- This subset is referred to as the *working set*, $\mathcal{W}_k$.
- It is required that the constraints in the working set be linearly independent.
- Let's assume all constraints are linearly independent.

$$\min \ \frac{1}{2}x^T Q x + q^T x, \quad \text{s.t.} \ \ a_i^T x \geq b_i, \ i \in \mathcal{I}$$

Algorithm: Active-Set Method for Convex QP

Choose a feasible x_0 and $\mathcal{W}_0 \leftarrow \mathcal{A}(x_0)$
for $k = 0, 1, \ldots$ **do**
  $p_k \leftarrow \arg\min \ \frac{1}{2} p^T Q p + (Qx_k + q)^T p$, s.t. $a_i^T p = 0, \ i \in \mathcal{W}_i$;
  if $p_k = 0$ **then**
    Find $\hat{\lambda}_i$ solving: $\sum_{i \in \mathcal{W}_k} a_i \lambda_i = Qx_k + q$;
    if $\lambda_i \geq 0, \forall i \in \mathcal{W}_k \cap \mathcal{I}$ **then**
      $x^\star \leftarrow x_k$; **stop**;
    else
      $j \leftarrow \arg\min_{j \in \mathcal{W}_k \cap \mathcal{I}} \ \lambda_j$; $\mathcal{W}_{k+1} \leftarrow \mathcal{W}_k \setminus \{j\}$;
    end if
  else ▷ $p_k \neq 0$
    $\mathcal{B} \leftarrow \arg\min_{i \notin \mathcal{W}_k, \ a_i^T p_k < 0} (b_i - a_i^T x_k)/(a_i^T p_k)$;
    $\alpha_k \leftarrow \min\left(1, (b_j - a_j^T x_k)/(a_j^T p_k)\right), \ j \in \mathcal{B}$;
    $x_{k+1} \leftarrow x_k + \alpha_k$; $\mathcal{W}_{k+1} \leftarrow \mathcal{W}_k \cup \{j\}_{j \in \mathcal{B}}$;
  end if
end for

Convergence in finite steps for $Q > 0$: in $\mathbf{P}_{\mathcal{X}}(x)$, $Q = I$, $q = 2x$.

Alternating Direction Method of Multipliers

- Consider the following problem

$$\begin{aligned} \min_{x,z} \quad & f_1(x) + f_2(z) \\ \text{s.t.} \quad & Ax = z \end{aligned}$$

- Define the augmented Lagrangian:

$$\mathbf{L}_A(x, z, \lambda; \mu) = f_1(x) + f_2(z) - \lambda^T (Ax - z) + \frac{\mu}{2} \|Ax - z\|^2$$

- The iterations become:

$$\begin{aligned} x_{k+1} &\in \arg\min_x \mathbf{L}_A(x, z_k, \lambda_k; \mu) \\ z_{k+1} &\in \arg\min_z \mathbf{L}_A(x_{k+1}, z, \lambda_k; \mu) \\ \lambda_{k+1} &= \lambda_k + \mu(Ax_{k+1} - z_{k+1}) \end{aligned}$$

- The inner two minimisations are decoupled.
- The method converges for convex f_1 and f_2 for any $\mu > 0$.
- It is related to the Augmented Lagrangian methods and Douglas-Raschford splitting.

Gradient Projection: ADMM

Algorithm: ADMM for QP with linear inequality constraints[a]

[a] For more detail see: Ghadimi, E., Teixeira, A., Shames, I. and Johansson, M., 2015. Optimal parameter selection for the alternating direction method of multipliers (ADMM): quadratic problems. IEEE Transactions on Automatic Control, 60(3), pp.644-658.

Choose x_0, z_0, u_0, and $\mu > 0$ ▷ u is the scaled Lagrange multiplier, $u = \lambda/\mu$

while A termination condition is not satisfied **do**

  $x_{k+1} \leftarrow -(Q + \rho A^\top A)^{-1}[q - \mu A^\top (z_k + u_k - c)]$;

  $z_{k+1} \leftarrow \max\{0, Ax_{k+1} - u_k - b\}$;

  $u_{k+1} \leftarrow u_k - Ax_{k+1} + b + z_{k+1}$;

end while

- The algorithm converges R-linearly to the solution for $Q > 0$: in $\mathbf{P}_{\mathcal{X}}(x)$, $Q = I$, $q = 2x$.
- Optimum step-length:

$$\mu^\star = \left(\sqrt{\lambda_{\max}(AQ^{-1}A^T)\lambda_{\min}(AQ^{-1}A^T)}\right)^{-1}$$

Projection: Approximating $\mathcal{X}$

- We know projection onto boxes is easy. So why not approximate constraints with a box?
- A box is a ∞-norm ball centred at x_c and with radius R:

$$\mathcal{B}(x_c, R) = \{x | \|x - x_c\|_\infty \leq R\}$$

- Let $\mathcal{X} = \{x | a_i^T x \leq b_i, i \in \mathcal{I}\}$.
- The goal is find the largest ball (square box) in $\mathcal{X}$.
- The problem of finding the largest ball (x_c is called the Chebyshev centre):

$$\begin{aligned} \max_{x_c, R} \quad & R \\ \text{s.t.} \quad & c_i(x_c, R) \leq 0, \quad i \in \mathcal{I} \\ & R \geq 0 \end{aligned}$$

$$\begin{aligned} c_i(x_c, R) &= \sup_{\|u\| \leq 1} a_i^T (x_c + Ru) - b_i = a_i^T x_c + R \left(\sup_{\|u\| \leq 1} a_i^T u \right) - b_i \\ &= a_i^T x_c + R \|a_i\|_* - b_i \quad (\|a_i\|_{\infty *} = \|a_i\|_1) \end{aligned}$$